

Spectrum Bandit Optimization

Marc Lelarge
INRIA – Ecole Normale Supérieure
email: marc.lelarge@ens.fr

Alexandre Proutiere and Sadegh Talebi
KTH Royal Institute of Technology
emails: alexandre.proutiere@ee.kth.se, mstms@kth.se

May 3, 2013

Abstract

We consider the problem of allocating radio channels to links in a wireless network. Links interact through interference, modelled as a conflict graph (i.e., two interfering links cannot be simultaneously active on the same channel). We aim at identifying the channel allocation maximizing the total network throughput over a finite time horizon. Should we know the average radio conditions on each channel and on each link, an optimal allocation would be obtained by solving an Integer Linear Program (ILP). When radio conditions are unknown a priori, we look for a sequential channel allocation policy that converges to the optimal allocation while minimizing on the way the throughput loss or *regret* due to the need for exploring sub-optimal allocations. We formulate this problem as a generic linear bandit problem, and analyze it first in a stochastic setting where radio conditions are driven by a stationary stochastic process, and then in an adversarial setting where radio conditions can evolve arbitrarily. We provide, in both settings, algorithms whose regret upper bounds outperform those of existing algorithms for linear bandit problems.

1 Introduction

Spectrum is a key and scarce resource in wireless communication systems, and it remains tightly controlled by regulation authorities. Most of the frequency bands are exclusively allocated to a single system licensed to use it everywhere and for periods of time that usually cover one or two decades. The consensus on this rigid spectrum management model is that it leads to significant inefficiencies in spectrum use. The explosion of demand for broadband wireless services also calls for more flexible models where much larger spectrum parts could be dynamically shared among users in a fluid manner. In such models, Dynamic Spectrum Access (DSA) techniques will play a major role. These techniques make it possible for radio devices to become frequency-agile, i.e. able to rapidly and dynamically access bands of a wide spectrum part.

In this paper, we consider wireless networks where transmitters can share a potentially large number of frequency bands or channels for transmission. In such networks, transmitters should be able to select a channel (i) that is not selected by neighbouring transmitters to avoid interference, and (ii) that offers good radio conditions. A spectrum allocation is defined by the channels assigned to the various transmitters or links, and our fundamental objective is to devise an optimal allocation, i.e., maximizing the network-wide throughput. If the radio conditions on each link and on each channel were known, the problem

would reduce to a combinatorial optimization problem, and more precisely to an Integer Linear Program. For example, if all links interfere each other (no two links can be active on the same channel), a case referred to as *full interference*, the optimal spectrum allocation problem is an instance of a Maximum Weighted Matching in a bipartite graph (vertices on one side correspond to links and vertices on the other side to channels; the weight of an edge, i.e., a (link, channel) pair, represents the radio conditions for the corresponding link and channel). In practice, the radio conditions on the various channels are not known a priori, and they evolve over time in an unpredictable manner. Hence, we need to dynamically learn and track the optimal spectrum allocation. This task is further complicated by the fact that we can gather information about the radio conditions for a particular (link, channel) pair only by actually including this pair in the selected spectrum allocation. We face a classical exploration vs. exploitation trade-off problem: we need to exploit the spectrum allocation with highest throughput observed so far whilst constantly exploring whether this allocation changes over time. We model our sequential spectrum allocation problem as a linear multi-armed bandit problem. The challenge in this problem resides in the very high dimension of the decision action space, i.e., in its combinatorial structure: the size of the set of possible allocations exponentially grows with the number of links and channels.

We study generic linear bandit problems in two different settings, and apply our results to sequential spectrum allocation problems. In the stochastic setting, we assume that the radio conditions for each (link,channel) pair evolve over time according to a stationary (actually i.i.d.) process whose mean is unknown. This first model is instrumental to represent scenarios where the average radio conditions evolve relatively slowly, in the sense that the spectrum allocation can be updated many times before this average exhibits significant changes. In the adversarial setting, the radio conditions evolve arbitrarily, as if they were generated by an *adversary*. This model is relevant when the channel allocation cannot be updated at the same pace as radio conditions change. In both settings, as usual for bandit optimization problems, we measure the performance of a given sequential decision policy through the notion of *regret*, defined as the difference of the performance obtained over some finite time horizon under the best static policy (i.e., assuming here that average radio conditions are known) and under the given policy. We make the following contributions:

- For stochastic linear bandit problems:
 - (a) we derive an asymptotic lower bound for the regret of any sequential decision policy, and show that this bound typically scales as $(n \times c) \log(T)$, where n , c , and T denote the number of links, the number of channels and the time horizon, respectively.
 - (b) We propose two sequential decision policies for linear bandit problems, that are simple extensions of the classical UCB and ϵ -greedy algorithms, and provide upper bounds on their regret, scaling respectively as $(n^4 \times c) \log(T)$ and $(n^2 \times c) \log(T)$. The latter constitutes the best regret upper bound for the linear bandit problem considered.
- For adversarial linear bandit problems: We propose ColorBand, a new sequential decision policy, and derive an upper bound on its regret. For example in the full interference case, when the numbers of channels and links are identical, this bound scales as $\sqrt{n^3 \log(n)T}$ (this improves over the upper bounds for the best previously known algorithms [6, 12], which scale as $\sqrt{n^5 \log(n)T}$).

Related work. Spectrum allocation has attracted considerable attention recently, mainly due to the increasing popularity of cognitive radio systems. In such systems, transmitters have to explore spectrum to find frequency bands free from primary users. This problem can also be formulated as a bandit problem, see e.g. [1, 13], but is simpler than our problem (in cognitive radio systems, there are basically c unknown variables, each representing the probability that a channel is free). Spectrum sharing problems similar to ours have been very recently investigated in [11, 16]. Both aforementioned papers restrict their analysis to the case of full interference, and even in this scenario, we obtain better regret bounds. As far

as we know, adversarial bandit problems have not been considered to model spectrum allocation issues. There is a vast literature on bandit problems, both in the stochastic and adversarial settings, see [5] for a quick survey. Surprisingly, there are very little work on linear bandit with discrete action space in the stochastic setting, and existing results are derived for very simple problems only, see e.g. [17] and references therein. In contrast, the problem has received more attention in the adversarial setting [4, 6, 9, 10, 12]. The algorithm we devise yields a regret upper bound that beats all known bounds of algorithms previously proposed in the literature.

2 Models and Objectives

2.1 Network and interference model

Consider a network consisting of n links indexed by $i \in [n] = \{1, \dots, n\}$. Each link can use one of the c available radio channels indexed by $j \in [c]$. Interference is represented as a conflict graph $G = (V, E)$ where vertices are links, and edges $(i, i') \in E$ if links i and i' interfere, i.e., these links cannot be simultaneously active. A spectrum allocation is represented as a configuration $M \in \{0, 1\}^{n \times c}$, where $M_{ij} = 1$ if and only if link- i transmitter uses channel j . M is feasible if (i) for all i , the corresponding transmitter uses at most one channel, i.e., $\sum_{j \in [c]} M_{ij} \in \{0, 1\}$; (ii) two interfering links cannot be active on the same channel, i.e., for all $i, i' \in [n]$, $(i, i') \in E$ implies for all $j \in [c]$, $M_{ij}M_{i'j} = 0$ ¹. Let \mathcal{M} be the set of feasible configurations. For $M \in \mathcal{M}$, if link i is active, we denote by $M(i)$ the channel allocated to this link. We also write $(i, j) \in M$ for $i \in [n]$ and $j \in [c]$, if link i is active under configuration M , and $j = M(i)$. In the following we denote by $\mathcal{K} = \{\mathcal{K}_\ell, \ell \in [k]\}$ the set of maximal cliques of the interference graph G . We also introduce $K_{\ell i} \in \{0, 1\}$ such that $K_{\ell i} = 1$ if and only if link i belongs to the maximal clique \mathcal{K}_ℓ .

We will pay a particular attention to the *full interference* case, where the conflict graph G is complete. In such a case, a feasible configuration M is just a matching in the complete bipartite graph $([n], [c])$, where on one side we have the set $[n]$ of links, and on the other side, the set $[c]$ of radio channels.

2.2 Fading

To model the way radio conditions evolve over time on the various channels, we consider a time slotted system, where the duration of a slot corresponds either to the transmission of a single packet or to that of a fixed number m of packets. The channel allocation, i.e., the chosen feasible configuration, may change at the beginning of each slot. We denote by $r_{ij}(t)$ the number of packets successfully transmitted during slot t when link- i transmitter selects channel j for transmission in this slot and in absence of interference. Depending on the ability of transmitters to switch channels, we introduce two settings.

In the *stochastic setting*, the number of successful packet transmissions $r_{ij}(t)$ on link i and channel j are independent over i and j , and are i.i.d. across slots t . The average number of successful packet transmission per slot is denoted by $\mathbb{E}[r_{ij}(t)] = \theta_{ij}$, and is supposed to be unknown initially. If m packets are sent per slot, $r_{ij}(t)$ is a random variable whose distribution is that of Y_{ij}/m where Y_{ij} has a binomial distribution $\text{Bin}(m, \theta_{ij})$. When $m = 1$, $r_{ij}(t)$ is a Bernoulli random variable of mean θ_{ij} . The stochastic setting models scenarios where the radio channel conditions are stationary.

In the *adversarial setting*, $r_{ij}(t) \in [0, 1]$ can be arbitrary (as if it was generated by an *adversary*), and unknown in advance. This setting is useful to model scenarios where the duration of a slot is comparable or smaller than the channel coherence time. In such scenarios, we assume that the channel allocation

¹This model assumes that the interference graph is the same over the various channels. Our analysis and results can be extended to the case where one has different interference graphs depending on the channel.

cannot change at the same pace as the radio conditions on the various links, which is of interest in practice, when the radios cannot rapidly change channels.

In the following, we denote by $r_M(t)$ the total number of packet successfully transmitted during slot t under feasible configuration $M \in \mathcal{M}$, i.e.,

$$r_M(t) = \sum_{i \in [n]} \sum_{j \in [c]} M_{ij} r_{ij}(t) = M \bullet r(t).$$

2.3 Channel allocations and objectives

We analyze the performance of adaptive spectrum allocation policies that may select different feasible configurations at the beginning of each slot, depending on the observed received throughput under the various configurations used in the past. More precisely, at the beginning of each slot t , under policy π , a feasible configuration $M^\pi(t) \in \mathcal{M}$ is selected. This selection is made based on some feedback on the previously selected configurations and their observed throughput. We consider two types of feedback.

Under *detailed feedback*, at the end of slot t , the number of packets successfully transmitted on the various links are observed, i.e., the feedback $f(t)$ is $(r_{ij}(t), i, j : M_{ij}^\pi(t) = 1)$. Under *aggregate feedback*, at the end of slot t , the total number of successfully sent packets is known, and so the feedback $f(t)$ is simply $r_{M^\pi(t)}(t)$. Aggregate feedback is of interest when we are not able to maintain the achieved throughput per link.

At the beginning of slot t , the selected configuration $M(t)$ may depend on past decisions and the received feedback, i.e., on $M^\pi(1), f(1), \dots, M^\pi(t-1), f(t-1)$. The chosen configuration can also be randomized (at the beginning of a slot, we sample a configuration from a given distribution that depends on past observations). We denote by Π the set of feasible policies. The objective is to identify a policy maximizing over a finite time horizon T the expected number of packets successfully transmitted or simply what we call the *reward*. The expectation is here taken with respect to the possible randomness in the stochastic rewards (in the stochastic setting) and in the probabilistic successively selected channel allocations. Equivalently, we aim at designing a sequential channel allocation policy that minimizes the *regret*. The regret of policy $\pi \in \Pi$ is defined by comparing the performance achieved under π to that of an idealised policy that assumes that the average conditions on the various links and channels are known:

$$R^\pi(T) = \max_{M \in \mathcal{M}} \mathbb{E} \left[\sum_{t=1}^T r_M(t) \right] - \mathbb{E} \left[\sum_{t=1}^T r_{M^\pi(t)}(t) \right], \quad (1)$$

where $M^\pi(t)$ denotes the action set selected in step t . The notion of regret quantifies the performance loss due to the need for learning radio channel conditions, and the above problem can be seen as a linear bandit problem.

3 Optimal Static Allocation

When evaluating the regret of a sequential spectrum allocation policy, the performance of the latter is compared to that of the best static allocation:

$$M^\star \in \arg \max_{M \in \mathcal{M}} \mathbb{E} \left[\sum_{t=1}^T r_M(t) \right],$$

where in the above formula, the expectation is taken with respect to the possible randomness in the throughput $r_M(t)$ (in the stochastic setting only). To simplify the presentation, we assume that the optimal static allocation M^\star is unique (the analysis can be readily extended to the case where several

configurations are optimal, but at the expense of the use of more involved notations). To identify M^* , we have to solve an Integer Linear Program (ILP). Let us first introduce the following set of ILPs parameterized by vector $r = (r_{ij}, i \in [n], j \in [c])$.

$$\begin{aligned}
& \max && \sum_{i \in [n], j \in [c]} r_{ij} M_{ij} \\
& \text{s.t.} && \sum_{j \in [c]} M_{ij} \leq 1, \quad \forall i \in [n], \\
& && \sum_{i \in [n]} K_{\ell i} M_{ij} \leq 1, \quad \forall \ell \in [k], j \in [c] \\
& && M_{ij} \in \{0, 1\}, \quad \forall i \in [n], j \in [c],
\end{aligned} \tag{2}$$

and denote by $V(r)$ its solution. In the stochastic setting, the performance of the best static policy is then $\mu^* = V(\theta)$, $\theta = (\theta_{ij}, i \in [n], j \in [c])$, whereas in the adversarial setting, the best static policy yields a reward equal to $V(\sum_{t=1}^T r(t))$.

Lemma 1 *The ILP problem (2) is NP-complete for general interference graphs.*

Indeed our ILP problem is a coloring problem of the interference graph G . If one considers all the links allocated to a given channel, we obtain a stable set of G . To be more precise, already with only one channel, our problem is NP-complete as when $c = 1$ and $r_{i1} = 1$ for all $i \in [n]$, then the optimum value of (2) is the stable set number of the interference graph G which is a NP-complete problem (Theorem 64.1 in [18]). It should be noticed that in contrast, when the interference graph is complete, i.e., in the full interference case, the ILP problem can be interpreted as a maximum weighted matching in a bipartite graph. As a consequence, it can be solved in polynomial time [18].

4 Stochastic Bandit Problem

This section is devoted to the analysis of our linear bandit problem in the stochastic setting. We first derive an asymptotic lower bound on the regret achieved by any feasible sequential spectrum allocation policy. This provides a fundamental performance limit that no policy can beat. We then present two policies that naturally extend strategies used in classical multi-armed bandit problems to linear bandit problems, and provide upper bounds on their respective regret. Most of the results presented here concern scenarios where detailed feedback is available.

4.1 Detailed feedback

4.1.1 Asymptotic regret lower bound

In their seminal paper [14], Lai and Robbins consider the classical multi-armed bandit problem, where a decision maker has to sequentially select an action from a finite set of K actions whose respective rewards are independent and i.i.d. across time. For example, when the rewards are distributed according to Bernoulli distributions of respective means

$\theta_1, \dots, \theta_K$, they show that the regret of any online action selection policy π satisfies the following lower bound:

$$\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq \sum_{i=1}^K \frac{\theta_1 - \theta_i}{KL(\theta_1, \theta_i)},$$

where without loss of generality $\theta_1 > \theta_i$ for all $i \neq 1$, and $KL(u, v)$ is the KL divergence number between two Bernoulli distributions of respective means u and v , $KL(u, v) = u \log(u/v) + (1 - u) \log(1 - u)/(1 - v)$. The simplicity of this lower bound is due to the stochastic independence of the rewards obtained selecting different actions. In our linear bandit problem, the rewards obtained selecting different configurations are inherently correlated (as in these configurations, a link may be allocated with the same channel). Correlations significantly complicate the derivation and the expression of the lower bound on regret. To derive such a bound, we use the techniques used in [8] to study the adaptive control of Markov chains.

We use the following notation: $\Theta = [0, 1]^{n \times c}$; $\theta = (\theta_{ij}, i \in [n], j \in [c])$; $\mu^M(\lambda) = M \bullet \lambda$, for any $M \in \mathcal{M}$ and $\lambda \in \Theta$. Recall that $\mu^* = \max_{M \in \mathcal{M}} M \bullet \theta$, and the optimal configuration is M^* , i.e., $\mu^* = M^* \bullet \theta$.

We introduce $B(\theta)$ as the set of *bad* parameters, i.e., the set of $\lambda \in \Theta$ such that configuration M^* provides the same reward as under parameter θ , and yet M^* is not the optimal static configuration:

$$B(\theta) = \{\lambda \in \Theta : (\forall i, j : M_{ij}^* = 1, \lambda_{ij} = \theta_{ij}), \text{ and } \mu^* < \max_{M \in \mathcal{M}} \mu^M(\lambda)\}.$$

Then $B(\theta) = \cup_{M \neq M^*} B_M(\theta)$ where

$$B_M(\theta) = \{\lambda \in \Theta : (\forall i, j : M_{ij}^* = 1, \lambda_{ij} = \theta_{ij}), \text{ and } \mu^* < \mu^M(\lambda)\}.$$

The reward distribution for link i under configuration M and parameter θ is denoted by $p_i(\cdot; M, \theta)$. This distribution is over the set $\mathcal{S} = \{0, 1\}$ when one packet is sent per slot, or the set $\mathcal{S} = \{0, 1/m, \dots, 1\}$ if m packets per slot are sent. Of course when $\sum_{j \in [c]} M_{ij} = 0$, we have $p_i(0; M, \theta) = 1$. When $\sum_{j \in [c]} M_{ij} = 1 = M_{iM(i)}$, if a single packet is sent per slot, we have, for $y_i \in \{0, 1\}$,

$$p_i(y_i; M\theta) = \theta_{iM(i)}^{y_i} (1 - \theta_{iM(i)})^{1-y_i},$$

and if m packet are sent, we have, for $y_i \in \{0, 1/m, \dots, 1\}$,

$$p_i(y_i; M, \theta) = \binom{m}{my_i} \theta_{iM(i)}^{my_i} (1 - \theta_{iM(i)})^{m-my_i}$$

We define the KL divergence number $KL^M(\theta, \lambda)$ under static configuration M as:

$$KL^M(\theta, \lambda) = \sum_{i \in [n]} \sum_{y_i \in \mathcal{S}} \log \frac{p_i(y_i; M, \theta)}{p_i(y_i; M, \lambda)} p_i(y_i; M, \theta).$$

For instance, when a single packet is sent per slot, we get:

$$KL^M(\theta, \lambda) = \sum_{i \in [n]} \sum_{j \in [c]} M_{ij} KL(\theta_{ij}, \lambda_{ij}).$$

As we shall see later in this section, we can identify sequential spectrum allocations whose regret scales as $\log(T)$ when T grows large. Hence we restrict our attention to so-called *uniformly good* policies: $\pi \in \Pi$ is uniformly good if for all $\theta \in \Theta$, if the configuration M is sub-optimal ($M \neq M^*$), then the number of times $T_M(t)$ it is selected up to time t satisfies: $\mathbb{E}[T_M(t)] = o(t^\gamma)$ for all $\gamma > 0$. We are now ready to state the regret lower bound.

Theorem 1 *For all $\theta \in \Theta$, for all uniformly good policy $\pi \in \Pi$,*

$$\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq C(\theta), \quad (3)$$

where $C(\theta)$ solves the following optimization problem:

$$\inf_{x_M \geq 0, M \in \mathcal{M}} \sum_{M \in \mathcal{M}} x_M (\mu^* - \mu^M(\theta)) \quad (4)$$

$$\text{s.t.} \quad \inf_{\lambda \in B_M(\theta)} \sum_{M \neq M^*} x_M KL^M(\theta, \lambda) \geq 1, \forall M \neq M^*. \quad (5)$$

The above lower bound is unfortunately not explicit. In the case of full interference, however, the bound can be simplified, and we may characterize how it scales with the numbers of links and channels.

Theorem 2 *In the case of full interference, when a single packet is sent per slot, we have*

$$C(\theta) = \Theta(n \times c), \quad \text{as } n, c \rightarrow \infty.$$

The above theorem states that there exist positive constants $k_1 > 0, k_2 > 0$ (that depend on θ) such that $C(\theta)/(nc) \in [k_1, k_2]$ for n, c large enough. This result is intuitive and means that the regret has to scale with the number of unknown parameters in the system.

4.1.2 Regret of UCB-like algorithms

We investigate here the performance of a variant of the celebrated UCB algorithm [2]. The idea of this variant is to attach to each (link, channel) pair an index q_{ij} that evolves over time, and to use these indexes to sequentially select configurations. A similar UCB-like algorithm has been recently proposed in [11].

Denote by $\hat{r}_{ij,s} = \frac{1}{s} \sum_{t=1}^s r_{ij}(t)$ the empirical average number of packets successfully sent over link i and channel j if channel j has been allocated s times to link i . Introduce $\rho_{s,t} = \sqrt{\frac{\alpha \log(t)}{s}}$. The index of the pair (link i , channel j) is defined by:

$$q_{ij}(t) = \hat{r}_{ij,T_{ij}(t)} + \rho_{T_{ij}(t),t},$$

where $T_{ij}(t)$ is the number of times channel j has been allocated to link i up to time t . The proposed variant of UCB algorithm selects at any slot the configuration maximizing the sum of the indexes of the (link, channel) pairs. To initialize the algorithm, we need to first build an index for each (link, channel) pair. To do so, we start by selecting a set $\mathcal{A} \subset \mathcal{M}$ of configurations that covers all possible pairs. The construction of such a set is easy, and for example, in the case of full interference, we can simply use a set of $\max(n, c)$ configurations or matchings. Let A be the cardinality of \mathcal{A} .

Algorithm 1: UCB Algorithm

Initialization: For $t = 1, \dots, A$, select configurations in \mathcal{A} , observe the detailed rewards, and update $q(t)$.
for all $t > A$ **do**
 Select configuration $M(t) \in \arg \max_{M \in \mathcal{M}} M \bullet q(t)$.
 Observe the detailed rewards, and update the vector of indexes $q(t+1)$.
end

Theorem 3 *For $\alpha > n + 1/2$, we have:*

$$R^{UCB}(T) \leq 4\alpha \frac{\Delta_{\max}}{\Delta_{\min}^2} n^3 c \log(T) + O(1), \quad \text{as } T \rightarrow \infty,$$

where $\Delta_{\max} = \max_{M \in \mathcal{M}} (\mu^* - \mu^M(\theta))$, $\Delta_{\min} = \min_{M \neq M^*} (\mu^* - \mu^M(\theta))$.

Note that this bound is valid for any interference graph. For the full interference case, however, it can be easily shown that the bound on the regret scales as $\frac{\Delta_{\max}^2}{\Delta_{\min}^2} nc \min(n, c)^3 \log(T)$. This is also consistent with the result derived in [11] where the analysis was restricted to the case $n \leq c$.

4.1.3 Regret of ϵ -greedy algorithms

ϵ -greedy algorithm consists in selecting the configuration that has provided with the maximum reward so far with probability $1 - \epsilon_t$, and a configuration selected uniformly at random among the covering set \mathcal{A} of configurations. By reducing the exploration rate ϵ_t over time, a logarithmic regret can be achieved. More precisely, we will choose $\epsilon_t = \min(1, d/t)$ for some constant $d > 0$. Define $\hat{r}(t) = (\hat{r}_{ij, T_{ij}(t)}, i \in [n], j \in [c])$.

Algorithm 2: ϵ -greedy Algorithm

Initialization: For $t = 1, \dots, A$, select configurations in \mathcal{A} , observe the detailed rewards, and update $\hat{r}_{ij, t}$.
for all $t > A$ **do**
 Let $\epsilon_t = \min(1, d/t)$.
 Select configuration $M(t) \in \arg \max_{M \in \mathcal{M}} M \bullet \hat{r}(t)$ with probability $1 - \epsilon_t$, and a configuration uniformly selected at random in \mathcal{A} with probability ϵ_t .
 Observe the detailed rewards and update $\hat{r}(t+1)$.
end

Theorem 4 *There exists a choice of parameter $d > 10An^2/\Delta_{\min}^2$ such that we have:*

$$R^{\epsilon\text{-greedy}}(T) \leq 10A \frac{\Delta_{\max}^2}{\Delta_{\min}^2} n^2 \log(T) + O(1) \quad \text{as } T \rightarrow \infty.$$

For the case of full interference, the bound on the regret can be improved to

$$R^{\epsilon\text{-greedy}}(T) \leq 10A \frac{\Delta_{\max}^2}{\Delta_{\min}^2} \min(n, c)^2 \log(T) + O(1).$$

Also notice that for this case, we can select \mathcal{A} with $A = \max(n, c)$. As a result, for the full interference case with this choice of A , the regret scales as $\frac{\Delta_{\max}^2}{\Delta_{\min}^2} \max(n, c) \min(n, c)^2 \log(T)$ when T grows large. Compared to the regret bound of UCB, a factor n^2 has been removed. The upper bound proposed in the above theorem is the best bound derived so far, even for the full interference case. Note however, that this does not imply that ϵ -greedy outperforms UCB algorithm, as we cannot compare algorithms by just looking at their respective regret upper bounds.

For the general interference graph, we can select \mathcal{A} with $A = \max(c, b)$ where b is the minimum number of channels required to obtain a feasible channel assignment (with respect to constraints of (2)) in which every node $i \in [n]$ is assigned a channel. It is obvious that feasible channel assignment over conflict graph G is equivalent to b -coloring problem of graph G which is to label vertices of G with b colors such that no two vertices sharing the same edge have the same color. As a result, we can select $b = \gamma(G)$, where $\gamma(G)$ is the chromatic number of conflict graph G , and finally $A = \max(c, \gamma(G))$. This also confirms the choice of $A = \max(c, n)$ for the full interference case as for complete graph G , $\gamma(G) = n$.

4.2 Aggregate feedback

We now briefly discuss the case where aggregate feedback only is available. We derive an asymptotic lower bound for regret in this scenario, but let for future work the design of sequential spectrum allocation strategies.

For any $M \in \mathcal{M}$, we define $\mathcal{A}_M = \{(i, j) \in [n] \times [c] : \sum_{l \in [c]} M_{il} = M_{ij} = 1\}$. Further introduce for all $k = 0, 1, \dots, n$:

$$p(k; M, \theta) = \sum_{A \subset \mathcal{A}_M, |A|=k} \prod_{(i,j) \in A} \theta_{ij} \prod_{(i,j) \in \mathcal{A}_M \setminus A} (1 - \theta_{ij}), \quad (6)$$

and

$$KL_2^M(\theta, \lambda) = \sum_{k=0}^n p(k; M, \theta) \log \frac{p(k; M, \theta)}{p(k; M, \lambda)}.$$

In the following theorem, we derive an asymptotic regret lower bound. This bound is different than that derived in Theorem 1, due to the different nature of the feedback considered. Comparing the two bounds may indicate the price to pay by restricting the set of spectrum allocation policies to those based on aggregate feedback only.

Theorem 5 *For all $\theta \in \Theta$, for all uniformly good policy $\pi \in \Pi$,*

$$\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq C_2(\theta), \quad (7)$$

where $C_2(\theta)$ solves the following optimization problem:

$$\inf_{x_M \geq 0, M \in \mathcal{M}} \sum_{M \in \mathcal{M}} x_M (\mu^* - \mu^M(\theta)), \quad (8)$$

$$s.t. \quad \inf_{\lambda \in B_M(\theta)} \sum_{M \neq M^*} x_M KL_2^M(\theta, \lambda) \geq 1, \forall M \neq M^*. \quad (9)$$

5 Adversarial Bandit Problem

In this section, we study the problem in the adversarial setting. In [3], a regret bound of $O(\sqrt{T})$ is derived in this setting, where the constant scales as the square root of the number of arms (up to logarithmic factors) and linearly with the reward of a maximal allocation. In our case, the number of arms typically grows exponentially with n even in simple cases. For example, in the full interference case, the number of possible allocations is the number of matching in the complete bipartite graph $([n], [c])$, i.e. $\frac{n!}{(n-c)!}$ if $n \geq c$. Also, in our case since $r_{ij}(t) \in [0, 1]$, the maximal reward of an allocation is of the order $\min(n, c)$. In the sequel, using the structure of our problem, we derive an algorithm with the same dependence in time as in [3] but with much lower constants, scaling as $n\sqrt{c \log c}$.

We start with some observations about the ILP problem (2):

$$\begin{aligned} \max_{M \in \mathcal{M}} r \bullet M &= \max_{p(M) \geq 0, \sum_{M \in \mathcal{M}} p(M) = 1} \sum_{M \in \mathcal{M}} p(M) r \bullet M \\ &= \max_{\mu \in Co(\mathcal{M})} r \bullet \mu, \end{aligned}$$

where $Co(\mathcal{M})$ is the convex hull of the feasible allocation matrices \mathcal{M} .

We identify matrices in $\mathbb{R}^{n \times c}$ with vectors in \mathbb{R}^{nc} . Without loss of generality, we can always assume that c is sufficiently large (possibly adding artificial channels with zero reward) such that for all $i \in [n]$,

$\sum_{j \in [c]} M_{ij} = n$ for all $M \in \mathcal{M}$, i.e. all links are allocated to a (possibly artificial) channel. Indeed, this can be done as soon as $c \geq \gamma(G)$ where $\gamma(G)$ is the chromatic number of the interference graph G . In other words, the bounds derived below are valid with c replaced by the maximum between the number of channels and the chromatic number of the interference graph. With this simplifying assumption, we can embed \mathcal{M} in the simplex of distributions in \mathbb{R}^{nc} by scaling all the entries by n . Let \mathcal{P} be this scaled version of $Co(\mathcal{M})$.

We also define the matrix in $\mathbb{R}^{n \times c}$ with coefficients $\mu_{ij}^0 = \frac{1}{n|\mathcal{M}|} \sum_{M \in \mathcal{M}} M_{ij}$. Clearly $\mu^0 \in \mathcal{P}$. We define $\mu_{\min} = \min n\mu_{ij}^0 \geq \frac{1}{|\mathcal{M}|}$. Our algorithms are inspired from [10] where full information is revealed and used the projection onto convex sets using the KL divergence (see Chapter 3, I-projections in [7]). We denote the KL divergence between distributions q and p in \mathcal{P} (or more generally in the simplex of distribution in \mathbb{R}^{nc}) by:

$$KL(q||p) = \sum_e q(e) \log \frac{q(e)}{p(e)},$$

where e ranges over the couples $(i, j) \in [n] \times [c]$ and with the usual convention where $p \log \frac{p}{q}$ is defined to be 0 if $p = 0$ and $+\infty$ if $p > q = 0$. By definition, the projection of a distribution q onto a closed convex set Ξ of distributions is the $p^* \in \Xi$ such that

$$KL(p^*||q) = \min_{p \in \Xi} KL(p||q). \quad (10)$$

5.1 Detailed feedback

We first present our algorithm in the case of detailed feedback.

Algorithm 3: ColorBand-1 Algorithm

Initialization: Start with the distribution $q_0 = \mu^0$.

for all $t \geq 1$ **do**

Let $p_{t-1} = (1 - \gamma)q_{t-1} + \gamma\mu^0$ ($p_{t-1} \in \mathcal{P}$ so that $np_{t-1} \in Co(\mathcal{M})$).

Select a random allocation $M(t)$ with distribution np_{t-1} .

Get a reward $r_t = \sum_{i,j} r_{ij}(t)M_{ij}(t)$ and observe the reward vector: $r_{ij}(t)$ for all ij such that $M_{ij}(t) = 1$.

Construct the reward matrix: $\tilde{r}_{ij}(t) = \frac{r_{ij}(t)}{np_{t-1}(ij)}$ for all i, j with $M_{ij}(t) = 1$ and all other entries are 0.

Update $\tilde{q}_t(ij) \propto q_{t-1}(ij) \exp(\eta \tilde{r}_{ij}(t))$.

Set q_t to be the projection of \tilde{q}_t onto the set \mathcal{P} using the KL divergence.

end

Theorem 6 *We have*

$$R^{ColorBand-1}(T) \leq 4n\sqrt{\mu_{\min}^{-1}T \log \mu_{\min}^{-1}}.$$

Note that in the full interference case, we have $\mu_{\min}^{-1} = \min(c, n)$.

Proof. We first prove the following result:

Lemma 2 We have for $\eta \leq 1$, and any $q \in \mathcal{P}$,

$$\sum_{t=1}^T q \bullet \tilde{r}(t) - \sum_{t=1}^T q_{t-1} \bullet \tilde{r}(t) \leq \eta \sum_{t=1}^T q_{t-1} \bullet \tilde{r}^2(t) + \frac{KL(q \| q_0)}{\eta},$$

where $\tilde{r}^2(t)$ is the vector that is the coordinate-wise square of $\tilde{r}(t)$.

Proof: We have

$$\begin{aligned} KL(q \| \tilde{q}_t) - KL(q \| q_{t-1}) &= \sum_e q(e) \log \frac{q_{t-1}(e)}{\tilde{q}_t(e)} \\ &= -\eta \sum_e q(e) \tilde{r}_e(t) + \log Z_t, \end{aligned}$$

with

$$\begin{aligned} \log Z_t &= \log \sum_e q_{t-1}(e) \exp(\eta \tilde{r}_e(t)) \\ &\leq \log \left(\sum_e q_{t-1}(e) (1 + \eta \tilde{r}_e(t) + \eta^2 \tilde{r}_e^2(t)) \right) \\ &\leq \eta q_{t-1} \bullet \tilde{r}(t) + \eta^2 \sum_e q_{t-1}(e) \tilde{r}_e^2(t), \end{aligned}$$

where we used $e^{\eta x} \leq 1 + \eta x + \eta^2 x^2$ for $\eta x \leq 1$ in the first inequality and $\log(1+x) \leq x$ for all $x > -1$ in the second inequality.

Hence, we have

$$KL(q \| \tilde{q}_t) - KL(q \| q_{t-1}) \leq -\eta q \bullet \tilde{r} + \eta q_{t-1} \bullet \tilde{r}(t) + \eta^2 q_{t-1} \bullet \tilde{r}^2(t).$$

Generalized Pythagorean inequality (see Theorem 3.1 in [7]) gives

$$KL(q \| q_t) + KL(q_t \| \tilde{q}_t) \leq KL(q \| \tilde{q}_t).$$

Since $KL(q_t \| \tilde{q}_t) \geq 0$, we get

$$KL(q \| q_t) - KL(q \| q_{t-1}) \leq -\eta q \bullet \tilde{r}(t) + \eta q_{t-1} \bullet \tilde{r}(t) + \eta^2 q_{t-1} \bullet \tilde{r}^2(t)$$

Summing over t gives

$$\sum_{t=1}^T (q \bullet \tilde{r}(t) - q_{t-1} \bullet \tilde{r}(t)) \leq \eta \sum_{t=1}^T q_{t-1} \bullet \tilde{r}^2(t) + \frac{KL(q \| q_0)}{\eta}.$$

■

Let \mathbb{E}_t be the expectation conditioned on all the randomness chosen by the algorithm up to time t . By construction, we have $\mathbb{E}_t[\tilde{r}_{ij}(t)] = r_{ij}(t)$, hence by linearity of expectation, we have $\mathbb{E}_t[q \bullet \tilde{r}(t)] = q \bullet r(t)$ and $\mathbb{E}_t[q_{t-1} \bullet \tilde{r}(t)] = q_{t-1} \bullet r(t)$. Moreover, we have

$$\begin{aligned} \mathbb{E}_t[q_{t-1} \bullet \tilde{r}^2(t)] &= \sum_{i \in [n]} \sum_{j \in [c]} q_{t-1}(ij) \frac{r_{ij}^2(t)}{n^2 p_{t-1}(ij)^2} n p_{t-1}(ij) \\ &= \sum_{i \in [n]} \sum_{j \in [c]} \frac{r_{ij}^2(t)}{n} \frac{q_{t-1}(ij)}{p_{t-1}(ij)} \\ &\leq \frac{c}{1-\gamma}, \end{aligned}$$

since $r_{ij}(t) \leq 1$ and $\frac{q_{t-1}(ij)}{p_{t-1}(ij)} \leq \frac{1}{1-\gamma}$.

Using Lemma 2 and the above bound, we get with nq^* the optimal allocation, i.e. $q^*(e) = \frac{1}{n}$ iff $M_e^* = 1$,

$$\begin{aligned} R^{ColorBand-1}(T) &= \mathbb{E} \left[\sum_{t=1}^T nq^* \bullet \tilde{r}(t) - \sum_{t=1}^T np_{t-1} \bullet \tilde{r}(t) \right] \\ &\leq \frac{ncT}{1-\gamma} + \frac{n \log \mu_{\min}^{-1}}{\eta} + n\gamma T, \end{aligned}$$

since $p_{t-1} \bullet \tilde{r}(t) - q_{t-1} \bullet \tilde{r}(t) \leq \gamma$ and

$$KL(q^* \| q_0) = -\frac{1}{n} \sum_{e \in M^*} \log n\mu_e^0 \leq \log \mu_{\min}^{-1}.$$

The proof is completed by setting $\eta = \sqrt{\frac{(1-\gamma) \log \mu_{\min}^{-1}}{cT}}$ and $\gamma = \sqrt{\frac{\mu_{\min}^{-1} \log \mu_{\min}^{-1}}{T}}$ which satisfy the necessary technical conditions. \square

5.2 Aggregate feedback

We now adapt our algorithm to deal with aggregate feedback.

Algorithm 4: ColorBand-2 Algorithm

Only steps 5 and 7 of ColorBand-1 algorithm are modified as follows:

5'. Get (and observe) a reward $r_t = \sum_{i,j} r_{ij}(t) M_{ij}(t)$.

7'. Let $\Sigma_{t-1} = \mathbb{E} [MM^T]$ where M has law np_{t-1} . Set $\tilde{r}(t) = r_t \Sigma_{t-1}^+ M(t)$, where Σ_{t-1}^+ is the pseudo-inverse of Σ_{t-1} .

Theorem 7 *We have*

$$R^{ColorBand-2}(T) = O \left(n \sqrt{\mu_{\min}^{-1} T \log \mu_{\min}^{-1}} \right).$$

Proof. We first prove a simple result:

Lemma 3 *For all $x \in \mathbb{R}^{nc}$, we have $\Sigma_{t-1}^+ \Sigma_{t-1} x = \bar{x}$, where \bar{x} is the orthogonal projection of x onto $\text{span}(\mathcal{M})$, the linear space spanned by \mathcal{M} .*

Proof: Note that for all $y \in \mathbb{R}^{nc}$, if $\Sigma_{t-1} y = 0$, then we have

$$y^T \Sigma_{t-1} y = \mathbb{E} [y^T M M^T y] = \mathbb{E} [(y^T M)^2] = 0, \quad (11)$$

where M has law $np_{t-1} = (1-\gamma)nq_{t-1} + \gamma n\mu^0$. By definition of μ^0 , each $M \in \mathcal{M}$ has a positive probability, so that by (11) $y^T M = 0$ for all $M \in \mathcal{M}$. In particular, we see that the linear application Σ_{t-1} restricted to $\text{span}(\mathcal{M})$ is invertible and is zero on $\text{span}(\mathcal{M})^\perp$, hence we have $\Sigma_{t-1}^+ \Sigma_{t-1} x = \bar{x}$. \blacksquare

Lemma 4 *We have*

$$\frac{\eta}{e^\eta - 1} \sum_{t=1}^T q \bullet \tilde{r}(t) - \frac{KL(q \| q_0)}{e^\eta - 1} \leq \sum_{t=1}^T q_{t-1} \bullet \tilde{r}(t).$$

Proof: The proof is the same as for Lemma 2 but in the upper bound for $\log Z_t$, we now use $e^{\eta x} \leq 1 + (e^\eta - 1)x$ valid for all $x \in [0, 1]$, the rest of the proof follows directly. \blacksquare

We have

$$\begin{aligned} \mathbb{E}_t [\tilde{r}(t)] &= \mathbb{E}_t [r_t \Sigma_{t-1}^+ M(t)] \\ &= \mathbb{E}_t [\Sigma_{t-1}^+ M(t) M(t)^T r(t)] \\ &= \Sigma_{t-1}^+ \Sigma_{t-1} r(t) = \overline{r(t)}, \end{aligned}$$

where the last equality follows from Lemma 3 and $\overline{r(t)}$ is the orthogonal projection of $r(t)$ onto $\text{span}(\mathcal{M})$. In particular, for any $np \in \text{Co}(\mathcal{M})$, we have

$$\mathbb{E} [np \bullet \tilde{r}(t)] = np \bullet \overline{r(t)} = np \bullet r(t).$$

To simplify notation, we denote $V_T = V(\sum_{t=1}^T r(t))$ and $\hat{V}_T = \sum_{t=1}^T r^{\text{ColorBand-2}}(t)$ the reward obtained by the algorithm ColorBand-2. Hence taking expectation in Lemma 4, we get

$$\mathbb{E} [\hat{V}_T] \geq (1 - \gamma) \left(\frac{\eta}{e^\eta - 1} \mathbb{E} [V_T] - \frac{n \log \mu_{\min}^{-1}}{e^\eta - 1} \right) - n\gamma T,$$

this gives

$$\begin{aligned} \mathbb{E} [V_T - \hat{V}_T] &\leq \left(1 - \frac{(1 - \gamma)\eta}{e^\eta - 1} \right) \mathbb{E} [V_T] \\ &\quad + (1 - \gamma) \frac{n \log \mu_{\min}^{-1}}{e^\eta - 1} + n\gamma T \\ &\leq \left(1 - \frac{(1 - \gamma)\eta}{e^\eta - 1} + \gamma \right) nT + \frac{n \log \mu_{\min}^{-1}}{e^\eta - 1}. \end{aligned}$$

Take $\eta = \gamma = \sqrt{\frac{\mu_{\min}^{-1} \log \mu_{\min}^{-1}}{T}}$ to get the result. \square

5.3 Implementation

There is a specific case where our algorithm can be efficiently implementable: when the convex hull $\text{Co}(\mathcal{M})$ can be captured by polynomial in n many constraints. Note that this cannot be ensured unless restrictive assumptions are made on the interference graph G since there are up to $3^{n/3}$ maximal cliques in a graph with n vertices [15]. There are families of graphs in which the number of cliques is polynomially bounded. These families include chordal graphs, complete graphs, triangle-free graphs, interval graphs, and planar graphs. Note however, that a limited number of cliques does not ensure a priori that $\text{Co}(\mathcal{M})$ can be captured by a limited number of constraints. To the best of our knowledge, this problem is open and only particular cases have been solved as for the stable set polytope (corresponding to the case $c = 2$, $r_{i1} = 1$ and $r_{i2} = 0$ with our notation) [18].

We consider the case where

$$\text{Co}(\mathcal{M}) = \left\{ \forall i, \sum_{j \in [c]} M_{ij} \leq 1, \quad \forall \ell, j, \sum_{i \in [n]} K_{\ell i} M_{ij} \leq 1 \right\}. \quad (12)$$

Note that in the special case where G is the complete graph, we have such a representation as in this case, we have

$$Co(\mathcal{M}) = \left\{ \sum_{j \in [c]} M_{ij} \leq 1, \quad \forall i, \sum_{i \in [n]} M_{ij} \leq 1, \quad \forall j \right\}.$$

We now give an algorithm for the step 6 of the algorithm, i.e. the projection onto \mathcal{P} . Since \mathcal{P} is a scaled version of $Co(\mathcal{M})$, we give an algorithm for the projection onto $Co(\mathcal{M})$ given by (12).

Set $\lambda_i(0) = \mu_j(0) = 0$ for all i, j and then define for $t \geq 0$,

$$\forall i \in [n], \lambda_i(t+1) = -\log \left(\sum_j M_{ij} e^{-\mu_j(t)} \right) \quad (13)$$

$$\forall j \in [c], \mu_j(t+1) = -\max_{\ell} \log \left(\sum_i K_{i\ell} M_{ij} e^{-\lambda_i(t+1)} \right). \quad (14)$$

We can show that

Proposition 1 *Let $M_{ij}^* = \lim_{t \rightarrow \infty} M_{ij} e^{-\lambda_i(t) - \mu_j(t)}$. Then M^* is the projection of M onto $Co(\mathcal{M})$ using the KL divergence.*

Although this algorithm is shown to converge, we must stress that the step (14) might be expensive as the number of distinct values of ℓ might be exponential in n . Again in the case of full interference, this step is easy and our algorithm reduces to Sinkhorn's algorithm (see [10] for a discussion).

Proof: First note that the definition of projection can be extended to non-negative vectors thanks to (10). More precisely, given an alphabet A and a vector $q \in \mathbb{R}_+^A$, we have for any probability vector $p \in \mathbb{R}_+^A$

$$\begin{aligned} \sum_{a \in A} p(a) \log \frac{p(a)}{q(a)} &\geq \sum_a p(a) \log \frac{\sum_a p(a)}{\sum_a q(a)} \\ &= \log \frac{1}{\|q\|_1}, \end{aligned}$$

thanks to the log-sum inequality. Hence we see that $p^*(a) = \frac{q(a)}{\|q\|_1}$ is the projection of q onto the simplex of \mathbb{R}_+^A .

Now define $\mathcal{A}_i = \{M_{ij}, \sum_j M_{ij} \leq 1\}$ and $\mathcal{B}_{\ell j} = \{M_{ij}, \sum_i K_{i\ell} M_{ij} \leq 1\}$. Hence $\cap_i \mathcal{A}_i \cap \cap_{\ell j} \mathcal{B}_{\ell j} = Co(\mathcal{M})$. By the argument described above, iteration (13) (resp. (14)) corresponds to the projection onto \mathcal{A}_i (resp. $\cap_{\ell} \mathcal{B}_{\ell j}$) and the proposition follows from Theorem 5.1 in [7]. ■

6 Conclusion

In this paper, we investigate the problem of sequential spectrum allocation in wireless networks where a potentially large number of channels are available, and whose radio conditions are initially unknown. The design of such allocations has been mapped into a generic linear multi-armed bandit problem, for which we have devised efficient online algorithms. Lower bounds for the performance of these algorithms have been derived, and they are shown to outperform performance bounds of existing algorithms, both in the stochastic setting where the radio conditions on the various channels and links are modelled as stationary

processes, and in the adversarial setting where no assumptions are made regarding the evolution of channel qualities. The practical implementation of our algorithms has just been briefly discussed. In particular, proposing efficient distributed implementations of these algorithms seems quite challenging, and we are currently working towards this objective.

References

- [1] A. Anandkumar, N. Michael, A. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *Selected Areas in Communications, IEEE Journal on*, 29(4):731–745, 2011.
- [2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- [3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77 (electronic), 2002/03.
- [4] B. Awerbuch and R. Kleinberg. Online linear optimization and adaptive routing. *J. Comput. Syst. Sci.*, 74(1):97–114, 2008.
- [5] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *to appear in Foundations and Trends in Machine Learning, available online*, 2012.
- [6] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 2012.
- [7] I. Csiszár and P. Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
- [8] T. L. Graves and T. L. Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM J. Control and Optimization*, 35(3):715–743, 1997.
- [9] A. Gyorgy, T. Linder, and G. Ottucsak. The shortest path problem under partial monitoring. In G. Lugosi and H. U. Simon, editors, *Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 468–482. Springer Berlin Heidelberg, 2006.
- [10] D. P. Helmbold and M. K. Warmuth. Learning permutations with exponential weights. *J. Mach. Learn. Res.*, 10:1705–1736, Dec. 2009.
- [11] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multi-player multi-armed bandits. *submitted to IEEE Trans. on Information Theory*, 2012.
- [12] S. Kale, L. Reyzin, and R. Schapire. Non-stochastic bandit slate problems. *Advances in Neural Information Processing Systems*, pages 1054–1062, 2010.
- [13] L. Lai, H. El Gamal, H. Jiang, and H. Poor. Cognitive medium access: Exploration, exploitation, and competition. *Mobile Computing, IEEE Transactions on*, 10(2):239–253, 2011.
- [14] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–2, 1985.
- [15] J. Moon and L. Moser. On cliques in graphs. *Israel Journal of Mathematics*, 3:23–28, 1965.

- [16] B. Radunovic, A. Proutiere, D. Gunawardena, and P. Key. Dynamic channel, rate selection and scheduling for white spaces. In *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*, CoNEXT '11. ACM, 2011.
- [17] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2), May 2010.
- [18] A. Schrijver. *Combinatorial Optimization : Polyhedra and Efficiency (Algorithms and Combinatorics)*. Springer, July 2004.

A Proof of Theorems 1 and 5

To derive regret lower bounds, we apply the techniques used by Graves and Lai [8] to investigate efficient adaptive decision rules in controlled Markov chains. We recall here their general framework. Consider a controlled Markov chain $(X_t)_{t \geq 0}$ on a finite state space \mathcal{S} with a control set U . The transition probabilities given control $u \in U$ are parameterized by θ taking values in a compact metric space Θ : the probability to move from state x to state y given the control u and the parameter θ is $p(x, y; u, \theta)$. The parameter θ is not known. The decision maker is provided with a finite set of stationary control laws $G = \{g_1, \dots, g_K\}$ where each control law g_j is a mapping from \mathcal{S} to U : when control law g_j is applied in state x , the applied control is $u = g_j(x)$. It is assumed that if the decision maker always selects the same control law g the Markov chain is then irreducible with stationary distribution π_θ^g . Now the reward obtained when applying control u leading to state x is denoted by $r(x, u)$, so that the expected reward achieved under control law g is: $\mu_\theta(g) = \sum_x r(x, g(x)) \pi_\theta^g(x)$. There is an optimal control law given θ whose expected reward is denoted $\mu_\theta^* \in \arg \max_{g \in G} \mu_\theta(g)$. Now the objective of the decision maker is to sequentially control laws so as to maximize the expected reward up to a given time horizon T . As for MAB problems, the performance of a decision scheme can be quantified through the notion of regret which compares the expected reward to that obtained by always applying the optimal control law.

Proof of Theorem 1. We now apply the above framework to our linear bandit problem. To simplify the presentation, we consider the case where in each slot, a single packet is transmitted. We will indicate what to modify when m packets are transmitted per slot.

The parameter θ takes values in $[0, 1]^{n \times c}$. The Markov chain has values in $\mathcal{S} = \{0, 1\}^n$. When m packets are transmitted per slot, $\mathcal{S} = \{0, 1/m, 2/m, \dots, 1\}^n$. The set of controls corresponds to the set of feasible configurations \mathcal{M} , and the set of control laws is also \mathcal{M} . These laws are constant, in the sense that the control applied by control law M does not depend on the state of the Markov chain, and corresponds to selecting configuration M . The transition probabilities are given as follows: for all $x, y \in \mathcal{S}$,

$$p(x, y; M, \theta) = p(y; M, \theta) = \prod_{i \in [n]} p_i(y_i; M, \theta),$$

where for all $i \in [n]$, if $\sum_{j \in [c]} M_{ij} = 0$, $p_i(0; M, \theta) = 1$, and if $\sum_{j \in [c]} M_{ij} = M_{iM(i)} = 1$, $p_i(y_i; M, \theta) = \theta_{iM(i)}^{y_i} (1 - \theta_{iM(i)})^{1-y_i}$. When m packets are sent per slot, the last formula has to be replaced by: $p_i(y_i; M, \theta) = \binom{m}{my_i} \theta_{iM(i)}^{my_i} (1 - \theta_{iM(i)})^{m-my_i}$. Finally, the reward $r(y, M)$ is defined by $r(y, M) = M \bullet y$. Note that the state space of the Markov chain is here finite, and so, we do not need to impose any cost associated with switching control laws (see the discussion on page 718 in [8]).

We can now apply Theorem 1 in [8]. Note that the KL number under configuration M is:

$$\begin{aligned} KL^M(\theta, \lambda) &= \sum_y \log \frac{p(y; M, \theta)}{p(y; M, \lambda)} p(y; M, \theta) \\ &= \sum_{i \in [n]} \sum_{y_i \in \{0,1\}} \log \frac{p_i(y_i; M, \theta)}{p_i(y_i; M, \lambda)} p_i(y_i; M, \theta). \end{aligned}$$

For instance, when a single packet is sent per slot, we get:

$$KL^M(\theta, \lambda) = \sum_{i \in [n]} \sum_{j \in [c]} M_{ij} KL(\theta_{ij}, \lambda_{ij}),$$

where $KL(u, v) = u \log(u/v) + (1-u) \log((1-u)/(1-v))$. From Theorem 1 in [8], we conclude that for any uniformly good rule π ,

$$\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq C(\theta),$$

where $C(\theta)$ is the optimal value of the following optimization problem:

$$\inf_{x_M \geq 0, M \in \mathcal{M}} \sum_{M \in \mathcal{M}} x_M (\mu^* - \mu^M(\theta)), \quad (15)$$

$$\text{s.t.} \quad \inf_{\lambda \in B(\theta)} \sum_{M \neq M^*} x_M KL^M(\theta, \lambda) \geq 1. \quad (16)$$

The result is obtained by observing that $B(\theta) = \bigcup_{M \neq M^*} B_M(\theta)$. □

Proof of Theorem 5. In the case of aggregate feedback, when configuration M is selected at time t , the global reward $r_M(t)$ only is known. To take this limited feedback into account, the state space of the corresponding Markov chain should record the global reward only. Hence, we have $\mathcal{S} = \{0, 1, \dots, n\}$. When the state is k , it means that the global received reward is equal to k . The probability that the reward under configuration M is equal to k is then $p(k; M, \theta)$ defined in (6), and so: for all $k', k \in \mathcal{S}$,

$$p(k', k; M, \theta) = p(k; M, \theta).$$

Theorem 5 is then a direct consequence of Theorem 1 in [8]. □

B Proof of Theorem 2

To simplify the presentation, we only consider the case where $n = c$. We first establish an upper bound on $C(\theta)$, and then derive a lower bound.

Recall that $C(\theta)$ solves the following optimization problem:

$$\inf_{x_M \geq 0, M \neq M^*} \sum_{M \neq M^*} x_M (\mu^* - \mu^M(\theta)) \quad (17)$$

$$\text{s.t.} \quad \inf_{\lambda \in B_M(\theta)} \sum_{Q \neq M^*} x_Q KL^Q(\theta, \lambda) \geq 1, \forall M \neq M^*. \quad (18)$$

B.1 Upper Bound for $C(\theta)$

We define \mathcal{L}_M as the set of links i such that $M(i) \neq M^*(i)$ and denote by L_M its cardinality. To obtain an upper bound for $C(\theta)$, we first derive a lower bound for $J_M^{(1)}(\theta), \forall M \neq M^*$, defined as:

$$\begin{aligned} J_M^{(1)}(\theta) &\triangleq \inf_{\lambda \in B_M(\theta)} \sum_{Q \neq M^*} x_Q KL^Q(\theta, \lambda) \\ &= \inf_{\lambda \in B_M(\theta)} \sum_{i \in \mathcal{L}_M} KL(\theta_{iM(i)}, \lambda_{iM(i)}) \sum_{k \in \mathcal{N}_i(M)} x_k, \end{aligned} \quad (19)$$

where $\mathcal{N}_i(M) = \{M' : (i, M(i)) \in M'\}$ is the set of configurations that share pair $(i, M(i))$ with M . Observe that the cardinality of $\mathcal{N}_i(M)$ is $(n-1)!$. Note that for two Bernoulli distributions with parameters $p, q \in (0, 1)$, we have $KL(p, q) \geq 2(p-q)^2$. Then

$$\begin{aligned} J_M^{(1)}(\theta) &\geq \inf_{\lambda \in \Theta} \sum_{i \in \mathcal{L}_M} 2(\theta_{iM(i)} - \lambda_{iM(i)})^2 \sum_{k \in \mathcal{N}_i(M)} x_k \\ &\quad \text{s.t.} \sum_{i \in \mathcal{L}_M} \lambda_{iM(i)} > \sum_{i \in \mathcal{L}_M} \theta_{iM^*(i)}, \\ &\quad \lambda_{iM^*(i)} = \theta_{iM^*(i)}, \quad \forall i \in [n]. \end{aligned} \quad (20)$$

To derive a lower bound for $J_M^{(1)}(\theta)$, we will use the following lemma, proved at the end of this section:

Lemma 5 *Let $\mathcal{A} \subseteq [n]$ be some set with cardinality A and $a \in \mathbb{R}_{++}^A$, $p \in \mathbb{R}_+^A$, and K be some constants such that $K \geq \sum_{i \in \mathcal{A}} p_i$. Define Z^* as*

$$Z^* \triangleq \inf \left\{ \sum_{i \in \mathcal{A}} 2a_i(p_i - z_i)^2 : \sum_{i \in \mathcal{A}} z_i > K, \quad z \in [0, 1]^A \right\}. \quad (21)$$

Then, $Z^* \geq \frac{2}{\sum_{i \in \mathcal{A}} \frac{1}{a_i}} (K - \sum_{i \in \mathcal{A}} p_i)^2$.

Applying the above lemma to problem (20), i.e., choosing for any i , $z_i = \lambda_{iM(i)}$, $a_i = \sum_{k \in \mathcal{N}_i(M)} x_k$, $p_i = \theta_{iM(i)}$, and $K = \sum_{i \in \mathcal{L}_M} \theta_{iM^*(i)}$, we obtain:

$$\begin{aligned} J_M^{(1)}(\theta) &\geq \frac{2 \left(\sum_{i \in \mathcal{L}_M} \theta_{iM^*(i)} - \sum_{i \in \mathcal{L}_M} \theta_{iM(i)} \right)^2}{\sum_{i \in \mathcal{L}_M} \frac{1}{\sum_{k \in \mathcal{N}_i(M)} x_k}} \\ &= \frac{2}{\sum_{i \in \mathcal{L}_M} \frac{1}{\sum_{k \in \mathcal{N}_i(M)} x_k}} (\Delta^M)^2. \end{aligned}$$

Now define:

$$\mathcal{D}(\theta) = \left\{ x \in \mathbb{R}_+^{|\mathcal{M}|-1} : \sum_{i \in \mathcal{L}_M} \frac{1}{\sum_{k \in \mathcal{N}_i(M)} x_k} \leq 2(\Delta^M)^2, \quad \forall M \neq M^* \right\}.$$

We have:

$$\mathcal{D}(\theta) \subseteq \left\{ x \in \mathbb{R}_+^{|\mathcal{M}|-1} : J_M^{(1)}(\theta) \geq 1, \quad \forall M \neq M^* \right\}.$$

Let $H(\theta) = \frac{n}{(n-1)!} \frac{1}{2\Delta_{\min}^2}$, and define the set $\mathcal{H}(\theta)$ as:

$$\mathcal{H}(\theta) \triangleq \left\{ x \in \mathbb{R}_+^{|\mathcal{M}|-1}, x_M \geq H(\theta) \right\}.$$

Then we have $\mathcal{H}(\theta) \subset \mathcal{D}(\theta)$. Indeed, if $x \in \mathcal{H}(\theta)$, for $M \neq M^*$,

$$\sum_{i \in \mathcal{L}_M} \frac{1}{\sum_{k \in \mathcal{N}_i(M)} x_k} \leq \frac{n}{H(\theta)(n-1)!} \leq 2(\Delta^M)^2.$$

We conclude that:

$$\begin{aligned} C(\theta) &\leq \inf_{x \in \mathcal{D}(\theta)} \sum_{M \neq M^*} x_M \Delta^M \leq \inf_{x \in \mathcal{H}(\theta)} \sum_{M \neq M^*} x_M \Delta^M \\ &\leq \Delta_{\max} \inf_{x \in \mathcal{H}(\theta)} \sum_{M \neq M^*} x_M = \frac{\Delta_{\max}}{2\Delta_{\min}^2} \cdot \frac{(n! - 1)n}{(n-1)!}. \end{aligned}$$

And thus: $C(\theta) \leq O(n^2)$ as $n \rightarrow \infty$.

B.2 Lower Bound for $C(\theta)$

To obtain the lower bound for $C(\theta)$, we derive an upper bound for $J_M^{(1)}(\theta)$. For any $(i, j) \notin M^*$, define $\xi_{ij} = \sum_{M \in \mathcal{L}_{(i,j)}} x_M$, where $\mathcal{L}_{(i,j)} = \{M : (i, j) \in M\}$ is the set of configurations including (i, j) . Then, for $M \neq M^*$, we have:

$$J_M^{(1)}(\theta) = \inf_{\lambda \in B_M(\theta)} \sum_{i \in \mathcal{L}_M} \xi_{iM(i)} KL(\theta_{iM(i)}, \lambda_{iM(i)}). \quad (22)$$

Let $B'_M(\theta) = \{\lambda \in B_M(\theta) : \lambda_{iM(i)} > \theta_{iM(i)}, \forall i \in \mathcal{L}_M\}$. Then,

$$J_M^{(1)}(\theta) \leq \inf_{\lambda \in B'_M(\theta)} \sum_{i \in \mathcal{L}_M} \xi_{iM(i)} KL(\theta_{iM(i)}, \lambda_{iM(i)}) \leq \underbrace{\inf_{\lambda \in B'_M(\theta)} \sum_{i \in \mathcal{L}_M} \xi_{iM(i)} (1 - \theta_{iM(i)}) \log \frac{1 - \theta_{iM(i)}}{1 - \lambda_{iM(i)}}}_{J_M^{(2)}(\theta)}, \quad (23)$$

where we used the fact that for two Bernoulli distributions with parameters $p, q \in (0, 1)$ with $p < q$, we have $KL(p, q) \leq (1 - p) \log \frac{1-p}{1-q}$. To derive an upper bound for $J_M^{(2)}(\theta)$, we will use the following result:

Lemma 6 *Let $\mathcal{A} \subseteq [n]$ be some set with cardinality A and $a, p \in \mathbb{R}_+^A$ and Q be some constants such that $\sum_{i \in \mathcal{A}} p_i \leq Q < A$. Define Y^* as*

$$\begin{aligned} Y^* &\triangleq \inf_{z \in [0,1]^A} \sum_{i \in \mathcal{A}} a_i (1 - p_i) \log \frac{1 - p_i}{1 - z_i} \\ &\quad s.t. \quad \sum_{i \in \mathcal{A}} z_i > Q, \\ &\quad z_i > p_i, \quad \forall i \in \mathcal{A}. \end{aligned} \quad (24)$$

Then, $Y^* \leq \frac{Q - \sum_{i \in \mathcal{A}} p_i}{A - Q} \sum_{i \in \mathcal{A}} a_i (1 - p_i)$.

Applying the previous lemma where $\mathcal{A} = \mathcal{L}_M$, $a_i = \xi_{iM^*}(i)$, $p_i = \theta_{iM}(i)$, $\forall i \in \mathcal{L}_M$, and $Q = \sum_{i \in \mathcal{L}_M} \theta_{iM^*}(i)$, we get:

$$\begin{aligned}
J_M^{(2)}(\theta) &\leq \frac{\sum_{i \in \mathcal{L}_M} \theta_{iM^*}(i) - \sum_{i \in \mathcal{L}_M} \theta_{iM}(i)}{L_M - \sum_{i \in \mathcal{L}_M} \theta_{iM^*}(i)} \sum_{i \in \mathcal{L}_M} (1 - \theta_{iM}(i)) \xi_{iM}(i) \\
&= \frac{\Delta^M}{L_M - \sum_{i \in \mathcal{L}_M} \theta_{iM^*}(i)} \sum_{i \in \mathcal{L}_M} (1 - \theta_{iM}(i)) \xi_{iM}(i) \\
&\leq \frac{\Delta^M \max_{i \in \mathcal{L}_M} (1 - \theta_{iM}(i))}{L_M - L_M \max_{i \in \mathcal{L}_M} \theta_{iM^*}(i)} \cdot \sum_{i \in \mathcal{L}_M} \xi_{iM}(i) \\
&\leq \frac{\Delta^M}{L_M} \cdot \frac{\max_{i \in \mathcal{L}_M} (1 - \theta_{iM}(i))}{\min_{i \in \mathcal{L}_M} (1 - \theta_{iM^*}(i))} \cdot \sum_{i \in \mathcal{L}_M} \xi_{iM}(i) \\
&\leq \frac{\beta \Delta_{\max}}{L_M} \sum_{i \in \mathcal{L}_M} \xi_{iM}(i),
\end{aligned}$$

where $\beta \triangleq \max_{M \neq M^*} \left(\frac{\max_{i \in \mathcal{L}_M} (1 - \theta_{iM}(i))}{\min_{i \in \mathcal{L}_M} (1 - \theta_{iM^*}(i))} \right)$. So far, we have established

$$J_M^{(1)}(\theta) \leq \frac{\beta \Delta_{\max}}{L_M} \sum_{i \in \mathcal{L}_M} \xi_{iM}(i).$$

Now, define

$$\mathcal{J}(\theta) = \left\{ x \in \mathbb{R}_+^{|\mathcal{M}|-1} : \frac{\beta \Delta_{\max}}{L_M} \sum_{i \in \mathcal{L}_M} \xi_{iM}(i) \geq 1, \quad \xi_{iM}(i) = \sum_{M' \in \mathcal{N}_i(M)} x_{M'}, \quad \forall M \neq M^*, \forall i \in \mathcal{L}_M \right\}.$$

Then: $\left\{ x \in \mathbb{R}_+^{|\mathcal{M}|-1} : J_M^{(1)}(\theta) \geq 1, \quad \forall M \neq M^* \right\} \subset \mathcal{J}(\theta)$, and hence:

$$C(\theta) \geq \inf_{x \in \mathcal{J}(\theta)} \Delta_{\min} \sum_{M \neq M^*} x_M. \quad (25)$$

Without loss of generality, from now on, we assume that $M^* = \{(i, i) : i \in [n]\}$. Define $\mathcal{T} = \{(1, j), j > 1\}$. We have:

$$\sum_M x_M = \xi_{11} + \sum_{j=2}^n \xi_{1j}.$$

Now since $\xi_{11} \geq x_{M^*}$:

$$\sum_{M \neq M^*} x_M \geq \sum_{j=2}^n \xi_{1j}.$$

Now, if

$$\Xi(\theta) = \left\{ (\xi_{ij})_{i,j \in [n], j \neq M^*(i)} \in \mathbb{R}_+^{n^2-n} : \sum_{i \in \mathcal{L}_M} \xi_{iM}(i) \geq \frac{L_M}{\beta \Delta_{\max}}, \quad \forall M \neq M^* \right\}, \quad (26)$$

using (25) and (26), we get:

$$C(\theta) \geq \inf_{x \in \mathcal{J}(\theta)} \Delta_{\min} \sum_{M \neq M^*} x_M \geq \inf_{\xi \in \Xi(\theta)} \Delta_{\min} \sum_{j=2}^n \xi_{1j}. \quad (27)$$

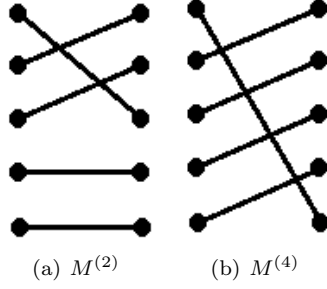


Figure 1: Two configurations $M^{(2)}$ and $M^{(4)}$ for the case of $n = 5$ and $M^* = \{(i, i) : i \in [n]\}$.

Next, we provide an explicit lower bound to the LP in the R.H.S. of (27). Consider $n-1$ configurations $M^{(k)}, k \in [n-1]$, where

$$M^{(k)} = \{(1, k+1), (2, 1), (3, 2), \dots, (k+1, k), (k+2, k+2), \dots, (n, n)\},$$

and verify that $\mathcal{T} \subset \bigcup_{k \in [n-1]} M^{(k)}$. Two examples of configurations, $M^{(2)}$ and $M^{(4)}$ for $n = 5$, are shown in Figure 2.

To obtain a lower bound for $\sum_{j=2}^n \xi_{1j} = \sum_{(i,j) \in \mathcal{T}} \xi_{ij}$, note that for $\xi \in \Xi(\theta)$:

$$\begin{aligned}
\xi_{12} &\geq \frac{2}{\beta \Delta_{\max}} - \xi_{21} \\
\xi_{13} &\geq \frac{3}{\beta \Delta_{\max}} - \xi_{21} - \xi_{32} \\
\xi_{14} &\geq \frac{4}{\beta \Delta_{\max}} - \xi_{21} - \xi_{32} - \xi_{43} \\
&\vdots \\
\xi_{1k} &\geq \frac{k}{\beta \Delta_{\max}} - \sum_{(i,j) \in M^{(k-1)} \setminus \{M^* \cup (1,k)\}} \xi_{ij} \\
&\vdots \\
\xi_{1n} &\geq \frac{n}{\beta \Delta_{\max}} - \sum_{(i,j) \in M^{(n-1)} \setminus \{(1,n)\}} \xi_{ij}.
\end{aligned}$$

Summing these inequalities and taking the infimum over $\xi \in \Xi(\theta)$, we get:

$$\inf_{\xi \in \Xi(\theta)} \sum_{(i,j) \in \mathcal{T}} \xi_{ij} \geq \frac{\beta}{\Delta_{\max}} \sum_{k=2}^n k - \inf_{\xi \in \Xi(\theta)} \sum_{(i,j) \in M^{(n-1)} \setminus \{(1,n)\}} w_{ij} \xi_{ij},$$

where $w_{ij}, (i, j) \in M^{(n-1)} \setminus \{(1, n)\}$ are strictly positive integers. We actually prove that:

$$\inf_{\xi \in \Xi(\theta)} \sum_{(i,j) \in M^{(n-1)} \setminus \{(1,n)\}} w_{ij} \xi_{ij} = 0.$$

For briefly, denote $Z = M^{(n-1)} \setminus \{(1, n)\}$. The above statement is equivalent to:

$$\exists \xi \in \Xi(\theta) : \forall (i, j) \in Z, \xi_{ij} = 0.$$

In view of the definition of $\Xi(\theta)$, this is then equivalent to showing that:

$$\forall M \neq M^*, \exists i \in \mathcal{L}_M : (i, M(i)) \notin Z,$$

which is easily verified. Indeed, let $i_0 = \min\{i : i \in \mathcal{L}_M\}$. Then $(i_0, M(i_0)) \notin Z$. Finally we have proved that:

$$\begin{aligned} C(\theta) &\geq \inf_{\xi \in \Xi(\theta)} \Delta_{\min} \sum_{(i,j) \in \mathcal{T}} \xi_{ij} \\ &\geq \frac{\Delta_{\min}}{\beta \Delta_{\max}} \sum_{k=2}^n k = \frac{\Delta_{\min}}{2\beta \Delta_{\max}} (n^2 + n - 1). \end{aligned}$$

This gives the required lower bound for $C(\theta)$. □

B.3 Supporting lemmas

Proof of Lemma 5. First observe that

$$Z^* \geq \min \left\{ \sum_{i \in \mathcal{A}} 2a_i(p_i - z_i)^2 : \sum_{i \in \mathcal{A}} z_i \geq K, \quad z \in \mathbb{R}_+^A \right\}.$$

The Lagrangian for this problem is given by

$$L(z, w) = \sum_{i \in \mathcal{A}} 2a_i(p_i - z_i)^2 + w(K - \sum_{i \in \mathcal{A}} z_i).$$

The minimizer of this problem z^* satisfies $\nabla_z L(z^*, w) = 0$. Hence, $0 = \frac{\partial L}{\partial z_i} \Big|_{z^*} = 4a_i(z_i^* - p_i) - w$ or equivalently $z_i^* = [p_i + \frac{w}{4a_i}]^+$. So the dual problem is

$$\begin{aligned} \max_{w \geq 0} \min_{z \in \mathbb{R}_+^A} L(z, w) &= \max_{w \geq 0} L(z^*, w) \\ &= \max_{w \geq 0} \sum_{i \in \mathcal{A}} 2a_i(p_i - z_i^*)^2 + w(K - \sum_{i \in \mathcal{A}} z_i^*) \\ &= \max_{w \geq 0} w^2 \sum_{i \in \mathcal{A}} \frac{1}{8a_i} + w \left(K - \sum_{i \in \mathcal{A}} \left(\frac{w}{4a_i} + p_i \right) \right) \\ &= \max_{w \geq 0} -w^2 \sum_{i \in \mathcal{A}} \frac{1}{8a_i} + w(K - \sum_{i \in \mathcal{A}} p_i) \end{aligned}$$

The maximizer of dual problem, w^* , solves

$$\begin{aligned} 0 &= \frac{d}{dw} \left[-w^2 \sum_{i \in \mathcal{A}} \frac{1}{8a_i} + w(K - \sum_{i \in \mathcal{A}} p_i) \right] \\ &= -2w \sum_{i \in \mathcal{A}} \frac{1}{8a_i} + (K - \sum_{i \in \mathcal{A}} p_i). \end{aligned}$$

Thus $w^* = \frac{4}{\sum_{i \in \mathcal{A}} \frac{1}{a_i}} (K - \sum_{i \in \mathcal{A}} p_i)$, and finally

$$Z^* \geq L(z^*, w^*) = \frac{2}{\sum_{i \in \mathcal{A}} \frac{1}{a_i}} (K - \sum_{i \in \mathcal{A}} p_i)^2.$$

□

Proof of Lemma 6. The partial Lagrangian for problem (24) is given by

$$L(z, w) = \sum_{i \in \mathcal{A}} a_i(1 - p_i) \log \frac{1 - p_i}{1 - z_i} + w(Q - \sum_{i \in \mathcal{A}} z_i).$$

The optimal solution to this problem z^* satisfies $\nabla_z L(z^*, w) = 0$ and $z_i^* \in [p_i, 1], \forall i \in \mathcal{A}$. Hence, $0 = \left. \frac{\partial L}{\partial z_i} \right|_{z^*} = \frac{a_i(1-p_i)}{1-z_i^*} - w$, which yields $z_i^* = \left[1 - \frac{a_i(1-p_i)}{w} \right]_{p_i}^1$ or equivalently $z_i^* = 1 - \frac{a_i(1-p_i)}{w}$, $w \geq \max_{i \in \mathcal{A}} a_i$. Thus, for $w \geq \max_{i \in \mathcal{A}} a_i$, the dual function is given by

$$\begin{aligned} D(w) &= \inf_{z \in \mathbb{R}_+^A} L(z, w) \\ &= \sum_{i \in \mathcal{A}} a_i(1 - p_i) \log \frac{w}{a_i} + w \left(Q - \sum_{i \in \mathcal{A}} \left[1 - \frac{a_i(1-p_i)}{w} \right] \right) \\ &= \sum_{i \in \mathcal{A}} a_i(1 - p_i) \log \frac{w}{a_i} + (Q - A)w + \sum_{i \in \mathcal{A}} a_i(1 - p_i) \end{aligned}$$

and the dual problem is: $\max_{w \geq \max_{i \in \mathcal{A}} a_i} D(w)$. Now, we establish an upper bound for Y^* , as follows

$$D(\tilde{w}) \triangleq \max_{w \geq 0} D(w) \geq \max_{w \geq \max_{i \in \mathcal{A}} a_i} D(w) = Y^*,$$

where the last equality is due to strong duality for problem (24). It can be easily shown that $\tilde{w} = \frac{\sum_{i \in \mathcal{A}} a_i(1-p_i)}{A-Q}$ and hence

$$\begin{aligned} D(\tilde{w}) &= \sum_{i \in \mathcal{A}} a_i(1 - p_i) \log \frac{\sum_{j \in \mathcal{A}} a_j(1-p_j)}{a_i(A-Q)} + (Q - A) \frac{\sum_{i \in \mathcal{A}} a_i(1-p_i)}{A-Q} + \sum_{i \in \mathcal{A}} a_i(1 - p_i) \\ &= \sum_{i \in \mathcal{A}} a_i(1 - p_i) \log \frac{\sum_{j \in \mathcal{A}} a_j(1-p_j)}{a_i(A-Q)}. \end{aligned}$$

Thus, we established

$$Y^* \leq \sum_{i \in \mathcal{A}} (1 - p_i) a_i \log \frac{\sum_{j \in \mathcal{A}} (1 - p_j) a_j}{a_i} - \left[\sum_{i \in \mathcal{A}} a_i(1 - p_i) \right] \log(A - Q). \quad (28)$$

In order to further simplify the R.H.S. of (28), we first derive an inequality using the log-sum inequality. Based on this inequality, for non-negative vectors e, f , we have

$$\sum_i e_i \log \frac{e_i}{f_i} \geq \sum_i e_i \log \frac{\sum_j e_j}{\sum_k f_k},$$

or equivalently, $\sum_i e_i \log \frac{\sum_k f_k}{f_i} \geq \sum_i e_i \log \frac{\sum_j e_j}{e_i}$. Now, choosing $e_i = (1 - p_i)a_i$ and $f_i = 1 - p_i, \forall i \in \mathcal{A}$, we get

$$\begin{aligned} & \sum_{i \in \mathcal{A}} (1 - p_i)a_i \log \frac{\sum_{k \in \mathcal{A}} (1 - p_k)}{1 - p_i} \geq \sum_{i \in \mathcal{A}} (1 - p_i)a_i \log \frac{\sum_{j \in \mathcal{A}} (1 - p_j)a_j}{(1 - p_i)a_i} \\ \Rightarrow & \left[\sum_{i \in \mathcal{A}} (1 - p_i)a_i \right] \log \sum_{k \in \mathcal{A}} (1 - p_k) \geq \sum_{i \in \mathcal{A}} (1 - p_i)a_i \log \frac{\sum_{j \in \mathcal{A}} (1 - p_j)a_j}{a_i}. \end{aligned}$$

Applying this result to the R.H.S. of (28), we then provide an upper bound to Y^* , as follows

$$\begin{aligned} Y^* & \leq \sum_{i \in \mathcal{A}} (1 - p_i)a_i \log \frac{\sum_{j \in \mathcal{A}} (1 - p_j)a_j}{a_i} - \left[\sum_{i \in \mathcal{A}} a_i(1 - p_i) \right] \log(A - Q) \\ & \leq \left[\sum_{i \in \mathcal{A}} a_i(1 - p_i) \right] \log \sum_{j \in \mathcal{A}} (1 - p_j) - \left[\sum_{i \in \mathcal{A}} a_i(1 - p_i) \right] \log(A - Q) \\ & = \log \frac{\sum_{i \in \mathcal{A}} (1 - p_i)}{A - Q} \sum_{i \in \mathcal{A}} a_i(1 - p_i) \\ & \leq \left(\frac{\sum_{i \in \mathcal{A}} (1 - p_i)}{A - Q} - 1 \right) \sum_{i \in \mathcal{A}} a_i(1 - p_i) \\ & = \frac{Q - \sum_{i \in \mathcal{A}} p_i}{A - Q} \sum_{i \in \mathcal{A}} a_i(1 - p_i), \end{aligned}$$

where in last two inequalities we used the fact that $\log z \leq z - 1$, for $z > 0$. This completes the proof. \square

C Proof of Theorem 3

The proof is along the same lines as in [11]. We present the proof for completeness. Let $T^M(t)$ denote the number of times spent on the suboptimal configuration $M \neq M^*$. Then the regret of UCB algorithm at time t is given by

$$R_{\text{UCB}}(t) = \sum_{M \neq M^*} \Delta^M \mathbb{E}[T^M(t)] \leq \Delta_{\max} \sum_{M \neq M^*} \mathbb{E}[T^M(t)].$$

For each suboptimal connections or (link, channel) pair (i, j) , define $Z_{ij}(t)$ as follows. If a suboptimal configuration $M(t)$ is chosen at time t , $Z_{ij}(t)$ is increased by one for some $(i, j) \in \arg \min_{(i', j') \in M(t)} T_{i'j'}(t)$, where $T_{i'j'}(t)$ is the number of times pair (i', j') has been selected up to time t . We let $\tilde{I}_{ij}(t)$ be the indicator function where $\tilde{I}_{ij}(t) = 1$ if $Z_{ij}(t)$ is incremented by 1 at time t . Therefore, when $\tilde{I}_{ij}(t) = 1$ there exists a configuration $M(t) \neq M^*$ such that $(i, j) \in M(t)$, where the time index in $M(t)$ implies that we may get different arms at different times.

Let $Z(t) = \sum_{i \in [n]} \sum_{j \in [c]} Z_{ij}(t)$. We have $\sum_{M \neq M^*} \mathbb{E}[T^M(t)] = \mathbb{E}[Z(t)]$. As a result, $R_{\text{UCB}}(t) \leq \Delta_{\max} \sum_{M \neq M^*} \sum_{i \in [n]} \mathbb{E}[Z_{iM(i)}(t)]$. Define $\bar{X}_{ij}(s) = \hat{r}_{ij, T_{ij}(s)}$ and

$$\mathcal{V}_M = \{i \in [n] : M_{ij} = 1 \text{ for some } j \in [c]\}$$

with cardinality V_M . Also recall that $\rho_{t,s} = \sqrt{\alpha \log t / s}$ and let u be a positive integer. Then, in (29) we provide an upper bound of $Z_{ij}(t)$, where for brevity we introduce $M_t = M(t) \neq M^*$ with $(i, j) \in M_t$.

$$\begin{aligned}
Z_{ij}(t) &\leq 1 + \sum_{s=A+1}^t \mathbb{1}\{\tilde{I}_{ij}(s) = 1\} \leq u + \sum_{s=u+1}^t \mathbb{1}\left\{\tilde{I}_{ij}(s) = 1, Z_{ij}(s-1) \geq u\right\} \\
&\leq u + \sum_{s=u+1}^t \mathbb{1}\left\{\sum_{i \in \mathcal{V}_{M^*}} \left(\bar{X}_{iM^*}(i)(s-1) + \rho_{s-1, T_{iM^*}(i)(s-1)}\right) \right. \\
&\quad \left. \leq \sum_{i \in \mathcal{V}_{M_s}} \left(\bar{X}_{iM_s}(i)(s-1) + \rho_{s-1, T_{iM_s}(i)(s-1)}\right), Z_{ij}(s-1) \geq u\right\} \\
&\leq u + \sum_{s=u+1}^t \mathbb{1}\left\{\min_{(0 < s_i < s), \forall i \in \mathcal{V}_{M^*}} \sum_{i \in \mathcal{V}_{M^*}} (\hat{r}_{iM^*}(i), s_i + \rho_{s-1, s_i}) \leq \max_{(u \leq s'_i < s), \forall i \in \mathcal{V}_{M_s}} \sum_{i \in \mathcal{V}_{M_s}} (\hat{r}_{iM_s}(i), s'_i + \rho_{s-1, s'_i})\right\} \\
&\leq u \\
&\quad + \sum_{s=1}^{\infty} \sum_{(1 \leq s_i \leq s-1), \forall i \in \mathcal{V}_{M^*}} \sum_{(u \leq s'_i \leq s-1), \forall i \in \mathcal{V}_{M_s}} \mathbb{1}\left\{\sum_{i \in \mathcal{V}_{M^*}} (\hat{r}_{iM^*}(i), s_i + \rho_{s, s_i}) \leq \sum_{i \in \mathcal{V}_{M_s}} (\hat{r}_{iM_s}(i), s'_i + \rho_{s, s'_i})\right\}.
\end{aligned} \tag{29}$$

Now the event

$$\left\{\sum_{i \in \mathcal{V}_{M^*}} (\hat{r}_{iM^*}(i), s_i + \rho_{s, s_i}) \leq \sum_{i \in \mathcal{V}_{M_s}} (\hat{r}_{iM_s}(i), s'_i + \rho_{s, s'_i})\right\}$$

implies at least one of the following events

$$\begin{aligned}
\mathcal{B}_{1,i} &\triangleq \left\{\hat{r}_{iM^*}(i), s_i \leq \theta_{iM^*}(i) - \rho_{s, s_i}\right\}, \quad \forall i \in \mathcal{V}_{M^*} \\
\mathcal{B}_{2,i} &\triangleq \left\{\hat{r}_{iM_s}(i), s'_i \geq \theta_{iM_s}(i) + \rho_{s, s'_i}\right\}, \quad \forall i \in \mathcal{V}_{M_s} \\
\mathcal{B}_3 &\triangleq \left\{\sum_{i \in \mathcal{V}_{M^*}} \theta_{iM^*}(i) \leq \sum_{i \in \mathcal{V}_{M_s}} \theta_{iM_s}(i) + 2 \sum_{i \in \mathcal{V}_{M_s}} \rho_{s, s'_i}\right\}.
\end{aligned}$$

Using Chernoff-Hoeffding bound, we easily get that $\Pr(\mathcal{B}_{1,i}) \leq s^{-2\alpha}, \forall i \in \mathcal{V}_{M^*}$ and $\Pr(\mathcal{B}_{2,i}) \leq s^{-2\alpha}, \forall i \in \mathcal{V}_{M_s}$. To obtain a bound on the probability of \mathcal{B}_3 , we use

$$u \geq \left\lceil \frac{4\alpha n^2 \log t}{\Delta_{\min}^2} \right\rceil.$$

From $s'_i \geq u$ and $V_{M'} \leq n, \forall M' \in \mathcal{M}$, we simply deduce that:

$$\sum_{i \in \mathcal{V}_{M^*}} \theta_{iM^*}(i) - \sum_{i \in \mathcal{V}_{M_s}} \theta_{iM_s}(i) - 2 \sum_{i \in \mathcal{V}_{M_s}} \rho_{s, s'_i} \geq \sum_{i \in \mathcal{V}_{M^*}} \theta_{iM^*}(i) - \sum_{i \in \mathcal{V}_{M_s}} \theta_{iM_s}(i) - \Delta_{\min} \geq 0.$$

Hence for $u \geq \left\lceil \frac{4\alpha n^2 \log t}{\Delta_{\min}^2} \right\rceil$, we get

$$\mathbb{E}[Z_{ij}(t)] \leq \frac{4\alpha n^2 \log t}{\Delta_{\min}^2} + \zeta_{\alpha} + 1$$

where $\zeta_\alpha \triangleq \sum_{t=1}^{\infty} (V_{M^*} + V_{M_t}) t^{V_{M^*} + V_{M_t} - 2\alpha}$. Noting that $V_{M'} \leq n, \forall M' \in \mathcal{M}$, the condition $\alpha > n + 1/2$ guarantees that $\zeta_\alpha < \infty$. Finally,

$$\mathbb{E}[Z(t)] \leq \frac{4\alpha n^3 c \log t}{\Delta_{\min}^2} + O(1), \quad \text{as } t \rightarrow \infty,$$

which completes the proof. \square

D Proof of Theorem 4

We first provide a bound on the probability of choosing a suboptimal configuration M . In what follows, we denote $X^M(t) = \sum_{i \in \mathcal{V}_M} \hat{r}_{iM(i), T_{iM(i)}(t)}$, $\Delta^M = \mu^* - \mu^M(\theta)$, where

$$\mathcal{V}_M = \{i \in [n] : M_{ij} = 1 \text{ for some } j \in [c]\}$$

with cardinality V_M .

For $t > d$, the probability of choosing a suboptimal configuration M can be written as

$$\Pr\{I_t = M\} \leq \frac{\epsilon_t}{A} \mathbb{1}\{M \in \mathcal{A}\} + (1 - \epsilon_t)E,$$

where $E = \Pr\{X^M(t-1) \geq X^{M^*}(t-1)\}$. Then

$$\begin{aligned} E &\leq \Pr\left\{\sum_{i \in \mathcal{V}_M} \hat{r}_{iM(i), T_{iM(i)}(t-1)} \geq \sum_{i \in \mathcal{V}_M} \left(\theta_{iM(i)} + \frac{\Delta^M}{2V_M}\right)\right\} \\ &+ \Pr\left\{\sum_{i \in \mathcal{V}_{M^*}} \hat{r}_{iM^*(i), T_{iM^*(i)}(t-1)} \leq \sum_{i \in \mathcal{V}_{M^*}} \left(\theta_{iM^*(i)} - \frac{\Delta^M}{2V_{M^*}}\right)\right\} \end{aligned}$$

Using the union bound we get

$$\Pr\left\{\sum_{i \in \mathcal{V}_M} \hat{r}_{iM(i), T_{iM(i)}(t-1)} \geq \sum_{i \in \mathcal{V}_M} \left(\theta_{iM(i)} + \frac{\Delta^M}{2V_M}\right)\right\} \leq \sum_{i \in \mathcal{V}_M} \underbrace{\Pr\left\{\hat{r}_{iM(i), T_{iM(i)}(t-1)} \geq \theta_{iM(i)} + \frac{\Delta^M}{2V_M}\right\}}_{\Pr\{\mathcal{B}_i(t-1)\}}$$

Now for $i \in \mathcal{V}_M$, using Chernoff-Hoeffding bound we get

$$\Pr\{\mathcal{B}_i(t)\} \leq \sum_{s=1}^t e^{-\frac{s}{2} \left(\frac{\Delta^M}{V_M}\right)^2} \Pr\left\{T_{iM(i)}(t) = s \mid \hat{r}_{iM(i), s} \geq \theta_{iM(i)} + \frac{\Delta^M}{2V_M}\right\} \quad (30)$$

Observe that for $w > 0$, $\sum_{s=x+1}^t e^{-ws} < \frac{1}{w} e^{-wx}$. This implies $\sum_{s=x+1}^t e^{-\frac{s}{2} \left(\frac{\Delta^M}{V_M}\right)^2} \leq 2 \left(\frac{V_M}{\Delta^M}\right)^2 e^{-\frac{x}{2} \left(\frac{\Delta^M}{V_M}\right)^2}$. Let $y_0 = \frac{1}{2A} \sum_{s=1}^t \epsilon_s$. We have:

$$\sum_{s=\lfloor y_0 \rfloor + 1}^t e^{-\frac{s}{2} \left(\frac{\Delta^M}{V_M}\right)^2} \Pr\left\{T_{iM(i)}(t) = s \mid \hat{r}_{iM(i), s} \geq \theta_{iM(i)} + \frac{\Delta^M}{2V_M}\right\} \leq 2 \left(\frac{V_M}{\Delta^M}\right)^2 e^{-\frac{\lfloor y_0 \rfloor}{2} \left(\frac{\Delta^M}{V_M}\right)^2}$$

We also have:

$$\begin{aligned}
& \sum_{s=1}^{\lfloor y_0 \rfloor} e^{-\frac{s}{2} \left(\frac{\Delta^M}{V_M} \right)^2} \Pr \left\{ T_{iM(i)}(t) = s \mid \hat{r}_{iM(i),s} \geq \theta_{iM(i)} + \frac{\Delta^M}{2V_M} \right\} \\
& \leq \sum_{s=1}^{\lfloor y_0 \rfloor} \Pr \left\{ T_{iM(i)}(t) = s \mid \hat{r}_{iM(i),s} \geq \theta_{iM(i)} + \frac{\Delta^M}{2V_M} \right\} \\
& \leq \sum_{s=1}^{\lfloor y_0 \rfloor} \Pr \left\{ T_{iM(i)}^R(t) \leq s \mid \hat{r}_{iM(i),s} \geq \theta_{iM(i)} + \frac{\Delta^M}{2V_M} \right\} \\
& \leq \sum_{s=1}^{\lfloor y_0 \rfloor} \Pr \left\{ T_{iM(i)}^R(t) \leq s \right\} \\
& \leq y_0 \Pr \left\{ T_{iM(i)}^R(t) \leq y_0 \right\},
\end{aligned}$$

where $T_{iM(i)}^R(t)$ denotes the number of times that pair $(i, M(i))$ is chosen during the exploration phase up to time t , and where we used the fact that $T_{iM(i)}^R(t) \leq T_{iM(i)}(t)$ implies $\Pr\{T_{iM(i)}(t) \leq y_0\} \leq \Pr\{T_{iM(i)}^R(t) \leq y_0\}$. Now it can be easily shown that: $\mathbb{E}[T_{iM(i)}^R(t)] = \text{Var}[T_{iM(i)}^R(t)] = \frac{1}{A} \sum_{s=1}^t \epsilon_s = 2y_0$. Using Bernstein's inequality, we have $\Pr\{T_{iM(i)}^R(t) \leq y_0\} \leq e^{-\frac{y_0}{5}}$. Finally, using $\epsilon_s = \min(1, d/s)$, we get $y_0 = \frac{d}{2A} + \frac{d}{2A} \log \frac{t}{d} = \frac{d}{2A} \log \frac{et}{d}$. In summary, we proved that:

$$\begin{aligned}
\Pr\{\mathcal{B}_i(t)\} & \leq y_0 e^{-\frac{y_0}{5}} + 2 \left(\frac{V_M}{\Delta^M} \right)^2 e^{-\frac{\lfloor y_0 \rfloor}{2} \left(\frac{\Delta^M}{V_M} \right)^2} \\
& = \frac{d}{2A} \log \frac{et}{d} \cdot e^{-\frac{d}{10A} \log \frac{et}{d}} + 2 \left(\frac{V_M}{\Delta^M} \right)^2 e^{-\frac{d}{4A} \left(\frac{\Delta^M}{V_M} \right)^2 \log \frac{et}{d}} \\
& = h(t) t^{-\frac{d}{10A}} + G_M t^{-\frac{d}{4A} \left(\frac{\Delta^M}{V_M} \right)^2}
\end{aligned}$$

where $h(t) \triangleq \frac{d}{2A} \left(\frac{e}{d} \right)^{-\frac{d}{10A}} \log \left(\frac{et}{d} \right)$ and $G_M \triangleq 2 \left(\frac{V_M}{\Delta^M} \right)^2 \left(\frac{e}{d} \right)^{-\frac{d}{4A} \left(\frac{\Delta^M}{V_M} \right)^2}$. As a result, we obtain:

$$\begin{aligned}
\Pr \left\{ \sum_{i \in \mathcal{V}_M} \hat{r}_{iM(i), T_{iM(i)}(t-1)} \geq \sum_{i \in \mathcal{V}_M} \left(\theta_{iM(i)} + \frac{\Delta^M}{2V_M} \right) \right\} & \leq \sum_{i \in \mathcal{V}_M} \Pr\{\mathcal{B}_i(t-1)\} \\
& \leq V_M h(t-1) (t-1)^{-\frac{d}{10A}} + V_M G_M \cdot (t-1)^{-\frac{d}{4A} \left(\frac{\Delta^M}{V_M} \right)^2}.
\end{aligned}$$

It can be shown similarly that

$$\begin{aligned}
& \Pr \left\{ \sum_{i \in \mathcal{V}_{M^*}} \hat{r}_{iM^*(i), T_{iM^*(i)}(t-1)} \leq \sum_{i \in \mathcal{V}_{M^*}} \left(\theta_{iM^*(i)} - \frac{\Delta^M}{2V_{M^*}} \right) \right\} \\
& \leq V_{M^*} h(t-1) (t-1)^{-\frac{d}{10A}} + V_{M^*} G_{M^*} (t-1)^{-\frac{d}{4A} \left(\frac{\Delta^M}{V_{M^*}} \right)^2},
\end{aligned}$$

where $G_{M^*} = 2 \left(\frac{V_{M^*}}{\Delta^M} \right)^2 \left(\frac{e}{d} \right)^{-\frac{d}{4A} \left(\frac{\Delta^M}{V_{M^*}} \right)^2}$.

We conclude that:

$$\begin{aligned} \Pr\{I_t = M\} &\leq \frac{d}{tA} \mathbb{1}\{M \in \mathcal{A}\} + (V_{M^*} + V_M) \left(1 - \frac{d}{t}\right) h(t-1)(t-1)^{-\frac{d}{10A}} \\ &\quad + V_M G_M \left(1 - \frac{d}{t}\right) (t-1)^{-\frac{d}{4A}} \left(\frac{\Delta^M}{V_M}\right)^2 + V_{M^*} G_{M^*} \left(1 - \frac{d}{t}\right) (t-1)^{-\frac{d}{4A}} \left(\frac{\Delta^M}{V_{M^*}}\right)^2 \end{aligned}$$

The upper bound on regret is then

$$\begin{aligned} R^{\epsilon-\text{greedy}}(t) &= \sum_{s=1}^t \sum_{M \neq M^*} \Delta^M \Pr\{I_s = M\} \\ &\leq \Delta_{\max} \sum_{s=1}^t \sum_{M \neq M^*} \Pr\{I_s = M\} \\ &\leq \Delta_{\max} \sum_{s=1}^t \sum_{M \neq M^*} \frac{d}{sA} \mathbb{1}\{M \in \mathcal{A}\} \\ &\quad + \Delta_{\max} \sum_{s=1}^t \sum_{M \neq M^*} (V_M + V_{M^*}) \left(1 - \frac{d}{s}\right) h(s-1)(s-1)^{-\frac{d}{10A}} \\ &\quad + \Delta_{\max} \sum_{s=1}^t \sum_{M \neq M^*} V_M G_M \left(1 - \frac{d}{s}\right) (s-1)^{-\frac{d}{4A}} \left(\frac{\Delta^M}{V_M}\right)^2 \\ &\quad + \Delta_{\max} \sum_{s=1}^t \sum_{M \neq M^*} V_{M^*} G_{M^*} \left(1 - \frac{d}{s}\right) (s-1)^{-\frac{d}{4A}} \left(\frac{\Delta^M}{V_{M^*}}\right)^2. \end{aligned} \tag{31}$$

Observe that $d > 10An^2/\Delta_{\min}^2$ implies $d > 10A$. Note further that $V_M \leq n, \forall M$. Then $d > 10An^2/\Delta_{\min}^2$ implies that $\forall M \neq M^*$

$$d > 4A \left(\frac{V_M}{\Delta^M}\right)^2 \quad \text{and} \quad d > 4A \left(\frac{V_{M^*}}{\Delta^M}\right)^2.$$

As a result, in the R.H.S. of (31), except the first term, the others will be bounded as t grows large. Then, after simplifications, we get:

$$R^{\epsilon-\text{greedy}}(t) \leq d\Delta_{\max} \log t + O(1), \quad \text{as } t \rightarrow \infty. \tag{32}$$

The proof is completed by taking the infimum in the R.H.S. over $d > 10An^2/\Delta_{\min}^2$.

□